A COMPOUND SEQUENTIAL BAYES PREDICTOR
FOR
SEQUENCES WITH AN EMPIRICAL MARKOV STRUCTURE

BY
THOMAS M. COVER
AARON SHENHAR

TECHNICAL REPORT NO. 11
OCTOBER 11, 1974

PREPARED UNDER THE AUSPICES
OF
AIR FORCE OFFICE OF SCIENTIFIC RESEARCH
CONTRACT AFOSR #F44620-74-C-0068

DEPARTMENT OF STATISTICS
STANFORD UNIVERSITY
STANFORD, CALIFORNIA
A Compound Sequential Bayes Predictor
for
Sequences with an Empirical Markov Structure

Thomas M. Cover†
Aaron Shenhar‡

Abstract
This paper is concerned with sequential predictors for binary sequences with no assumptions upon the existence of an underlying process. The rule offered here induces an expected proportion of errors which differs by $O(n^{-1/2})$ from the Bayes envelope with respect to the observed $k^{th}$ order Markov structure. This extends the compound sequential Bayes work of Robbins, Hannan and Blackwell from sequences with perceived $0^{th}$ order structure to sequences with perceived $k^{th}$ order structure. The proof follows immediately from applying the $0^{th}$ order theory to $2^k$ separate subsequences.

†Stanford University, Stanford, California
Submitted September, 1974 to the Annals of Statistics.
This work was supported by contract AFOSR #F44620-74-C-0068.
1. Introduction

We are interested in sequential prediction procedures that exploit any apparent order in the sequence. We do not assume the existence of any underlying statistics, but assume that the sequence is an outcome of a game against a malevolent intelligent nature. We shall show that if the sequence has any $k$th order Markov structure, the expected sequential prediction score will be as high as if the predictor had known these empirical statistics at the beginning. Thus nature cannot "set up" the predictor for future disastrous predictions.

Let $x \in \{0,1\}^\infty$. At time $n$ we observe the sequence $x(n) = (x_1, x_2, \ldots, x_n) \in \{0,1\}^n$. How do we predict $x_{n+1}$? We shall base our guess on the observed empirical distribution of sequences of length $k + 1$ which have the first $k$ bits equal to $(x_{n-k+1}, \ldots, x_n)$. This approach divides the question into $2^k$ parallel problems indexed by the immediate past.

Let $\delta(n) = (\delta_1, \delta_2, \ldots, \delta_n) \in \{0,1\}^n$ be a sequence of guesses for $x_1, x_2, \ldots, x_n$. Then the total prediction score (the fraction of correct guesses) will be given by

$$S_n = \frac{1}{n} \sum_{i=1}^{n} \left( 1 - |\delta_i - x_i| \right).$$

(1)

Let a random sequential predictor be described by $P = (p_1, p_2, \ldots, p_n)$, with the interpretation that we predict $x_i = 1$ with probability $p_i = p_i(x_1, x_2, \ldots, x_{i-1}, \delta_1, \delta_2, \ldots, \delta_{i-1}), 0 \leq p_i \leq 1$. The expected empirical average score will be given by

$$\bar{S}_n = \frac{1}{n} \sum_{i=1}^{n} \left( p_i x_i + (1-p_i)(1-x_i) \right).$$

(2)
Robbins [1], Hannan [2], and Blackwell [3] and [4] have considered algorithms for the $0^{th}$ order prediction problems which achieve $\hat{S}_n$ near the "Bayes envelope"

$$\hat{S}_n = \max \left\{ \frac{1}{n} \sum_{i=1}^{n} x_i, 1 - \frac{1}{n} \sum_{i=1}^{n} x_i \right\} \quad (3)$$

with an error bounded by $c/\sqrt{n}$. See also Cover [5] for the optimal solution when $n$ is known in advance. Swain [6] and Johns [7] have considered some other compound sequential decision problems in which there is a sequence of associated observations of the states of nature, and the expected score is compared to the $k^{th}$ order empirical distribution. See also Tainiter [8] for an empirical Bayes prediction procedure assuming there exists a true underlying source of known order.

In the next section we present a natural goal for a prediction process when a sequence seems to have some Markov structure. In section 3 we review Blackwell's algorithm for a prediction procedure for the zero order problem based on a two person zero sum game. In section 4 we then apply this algorithm to derive an expression for the asymptotic behavior of the expected score with respect to the Bayes envelope for the empirical statistics that arise from the $k^{th}$ order Markov structure.

2. A natural goal for the prediction score

Suppose that a sequence $x(n) \in \{0,1\}^n$ is observed that seems to have a Markov structure, i.e., the empirical Markov matrix of some order is far from Bernoulli. Clearly a statistician would be unhappy to have incurred a prediction score greatly worse than the Bayes risk with respect to this
perceived structure. The natural goal we wish to achieve is described as follows. For a moment, suppose we have a true $k^{th}$ order ergodic Markov process $\{X_i\}_{i=1}^{\infty}, X_i \in \{0,1\}$, with known statistics. Let a state of the process be denoted by $z \in \{0,1\}^k$ and let the transition probability matrix be $\bar{P}(z'|z), z, z' \in \{0,1\}^k$. For each $z$ let the nonzero elements of the transition probability matrix be $P(1|z)$ and $P(0|z) = 1 - P(1|z)$, the respective probabilities that 1 and 0 follow state $z$. Let $\mu(z)$ be the stationary probabilities that follow state $z$. Let $\mu(z)$ be the stationary distribution on the state space.

A Bayes predictor will induce the following steady state probability of a correct guess:

$$P^*_c(\mu, P) = \sum_{z \in \{0,1\}^k} \mu(z) \max \left\{ P(1|z), 1 - P(1|z) \right\}.$$ (4)

Returning to the nonstatistical problem, it follows that a desirable score for our observed sequence will be $S_n = P^*_c(\hat{\mu}, \hat{P})$, where $\hat{\mu}, \hat{P}$ are the empirical statistics induced by $x(n)$ as follows. Let $z, 1$ and $z, 0$ be the sequence $z$ followed by 1 and 0 respectively. Let $n(z, 1)$ and $n(z, 0)$ be respectively the number of times the sequences $z, 1$ and $z, 0$ were observed in $x(n)$ and let $n'(z)$ be the number of times the sequence $z$ was observed in $x(n-1) = (x_1, x_2, \ldots, x_{n-1})$. We observe that

$$n'(z) = n(z, 1) + n(z, 0).$$ (5)

Now define

$$\hat{P}(1|z) = \frac{n(z, 1)}{n'(z)}, \quad \hat{P}(0|1) = \frac{n(z, 0)}{n'(z)}$$ (6)
\[ \hat{\mu}(z) = \frac{n'(z)}{\sum_{z \in \{0,1\}^k} n'(z)}. \] (6 cont.)

Then \( \hat{P}(1|z) \) and \( \hat{P}(0|z) \) define transition probabilities for a \( k \)-th order Markov process with a stationary distribution \( \hat{\mu}(z) \).

3. A random predictor with \( k = 0 \) (Blackwell's procedure)

Let \( \mathbb{E}^2 \) be two dimensional Euclidean space with coordinates \( \bar{x}_n, S_n \), where \( \bar{x}_n = \frac{1}{n} \sum_{i=1}^{n} x_i \), and \( S_n \) is the empirical average score given by (1). Let the point \((\bar{x}_n, S_n) \in \mathbb{E}^2\) be the current result of the game so far. On the next step, we predict \( x_{n+1} = 1 \) and \( x_{n+1} = 0 \) with probabilities \( p_{n+1} \) and \( 1 - p_{n+1} \) respectively. Following Blackwell's definition of approachability of sets, the convex set in \( \mathbb{E}^2 \) defined by the Bayes envelope \( \hat{S}_n \) is achievable with the following procedure:

\[
p_{n+1} = \begin{cases} 
0, & \text{for } |\bar{x}_n - \frac{1}{2}| > |S_n - \frac{1}{2}| \text{ and } \bar{x}_n \leq \frac{1}{2} \\
1, & \text{for } |\bar{x}_n - \frac{1}{2}| > |S_n - \frac{1}{2}| \text{ and } \bar{x}_n > \frac{1}{2} \\
.5, & \text{for } |\bar{x}_n - \frac{1}{2}| < |S_n - \frac{1}{2}| \text{ and } S_n > \frac{1}{2} \\
\frac{\bar{x}_n - S_n}{1 - 2S_n}, & \text{for } |\bar{x}_n - \frac{1}{2}| < |S_n - \frac{1}{2}| \text{ and } S_n \leq \frac{1}{2}
\end{cases}
\quad (7)

Hannan [2, p-139] has indicated that the Blackwell procedure satisfies

\[ \hat{S}_n - \bar{S}_n \preceq \frac{3}{\sqrt{n}}, \quad \forall n, \forall x \in [0,1]^\infty, \] (8)
i.e., one can get arbitrarily close to the Bayes envelope uniformly in $x(n)$.

The extension of these results to $k^{th}$ order dependence is found by applying Blackwell's procedure to a series of parallel problems determined by the current state of the observed sequence.

4. $k^{th}$ order Markov predictor

At time $n$, $n = k + 1, \ldots$, let $z = (x_{n-k+1}, \ldots, x_n)$ be the current state of the observed sequence. We now predict the next bit according to the distribution of the sequences $z, 1$ and $z, 0$ in $x_1, x_2, \ldots, x_n$, using the procedure described in section 3. Therefore for each $z$ we have a separate prediction game and each game is played exactly $n'(z)$ times.

Let $\bar{S}_n'(z)$ be the expected score of the game indexed by state $z$, and let the partial Bayes envelope for the $z$ game be denoted by

$$\hat{S}_n'(z) = \max \left\{ \frac{n(z, 1)}{n'(z)}, \frac{n(z, 0)}{n'(z)} \right\} = \max \left\{ \hat{P}(1|z), 1 - \hat{P}(1|z) \right\}. \quad (9)$$

We can immediately apply Blackwell's bound to each game. Thus $\bar{S}_n'(z)$ will satisfy

$$\hat{S}_n'(z) - \bar{S}_n'(z) \leq \frac{3}{\sqrt{n'(z)}} \quad (10)$$

$$\forall z \in \{0, 1\}^k, \forall n'(z), \forall x \in [0, 1]^\infty.$$ 

Let $m$ be the total number of prediction plays. Using equation (5) for $n'(z)$ we obtain
\[ m = \sum_{z \in \{0,1\}^k} n'(z) = n - k \]  \hspace{1cm} (11)

(The predictor is not defined on the first \( k \) terms of \( x(n) \).) Now the average expected score over \( m = n - k \) plays will be

\[ \bar{s}_m = \frac{1}{m} \sum_{z \in \{0,1\}^k} n'(z) \bar{s}_{n'}(z) \]  \hspace{1cm} (12)

However we are interested in the total score achieved over the full length of \( x(n) \). Therefore, counting no score for the first \( k \) plays,

\[ \bar{s}_n = \frac{1}{n} \sum_{z \in \{0,1\}^k} n'(z) \bar{s}_{n'}(z) = \frac{n-k}{n} \bar{s}_m \]  \hspace{1cm} (13)

But the total \( k^{th} \) order Bayes envelope is

\[
P_c^*(n, \hat{\mu}, \hat{p}) = \sum_{z \in \{0,1\}^k} \hat{\mu}(z) \max\left\{\hat{p}(1|z), 1 - \hat{p}(1|z)\right\}
\]

\[ = \frac{1}{n-k} \sum_{z \in \{0,1\}^k} n'(z) \bar{s}_{n'}(z) \]

where \( \hat{\mu}(z) \) and \( \hat{p}(1|z) \) are the empirical statistics given by (6). We can therefore establish the following.

Theorem:

The prediction procedure in equation (7), performed conditionally on each \( z \in \{0,1\}^k \), achieves
\[ P^*_c(n, \hat{\mu}, \hat{p}) - \bar{S}_n \leq \frac{2^k 3}{\sqrt{n}} + \frac{k}{n}, \quad \forall x \in \{0,1\}^\infty, \forall n. \quad (15) \]

Thus the expected score \( \bar{S}_n \) is as high as if we had initially known the empirical \( k^{th} \) order Markov structure.

Proof:

\[ P^*_c(n, \hat{\mu}, \hat{p}) - \bar{S}_n = \frac{1}{n-k} \sum_{z \in \{0,1\}} k n'(z) \hat{S}'_n(z) - \frac{1}{n} \sum_{z \in \{0,1\}} k n'(z) \bar{S}'_n(z) \]

\[ = \frac{1}{n} \sum_{z \in \{0,1\}} k n'(z) \left( \hat{S}'_n(z) - \bar{S}'_n(z) \right) + \frac{k}{n(n-k)} \sum_{z \in \{0,1\}} k n'(z) \hat{S}'_n(z). \quad (16) \]

From (10) and the fact \( \frac{1}{n-k} \sum_{z \in \{0,1\}} k n'(z) \hat{S}'_n(z) \leq 1 \), we obtain

\[ P^*_c(n, \hat{\mu}, \hat{p}) - \bar{S}_n \leq \frac{1}{n} \sum_{z \in \{0,1\}} k n'(z) \frac{3}{\sqrt{n}(z)} + \frac{k}{n} \]

\[ \leq \frac{2^k 3}{n} \sqrt{\max_{z \in \{0,1\}} n'(z)} + \frac{k}{n} \leq \frac{2^k 3}{\sqrt{n}} + \frac{k}{n}. \quad (17) \]

5. Remarks

Although our motivation is nonstatistical and not tied to the existence of a true underlying process, we remark that if \( \{X_i\}_{i=1}^\infty \) is indeed a \( k^{th} \) order Markov process, the predictor given in the theorem is asymptotically Bayes in the sense that
\[ \overline{s} \to P^*_c (\mu, P) \text{ wp 1.} \] (18)

If \( \{X_i\}_{i=1}^\infty \) is a sample sequence from a \( k^{th} \) order Markov process with unknown statistics, it can be seen that there exists a deterministic sequential predictor which learns the statistics and asymptotically achieves the Bayes risk. One may wonder why randomization is required for the sequential predictor studied here. The answer is essentially game theoretic. In fact it can be easily seen (Cover [5]) that for any deterministic sequential predictor there exists a sequence for which 

\[ \overline{s}(x(n)) = 0 \quad \text{and} \quad P^*_c(n, \hat{\mu}, \hat{P}) = \frac{1}{2}. \]

The Blackwell predictor could be replaced by other asymptotically compound sequential Bayes predictors. For example, using Hannan's procedure [2, section 7] for a Bayes predictor, where uniformly distributed random variables weighted by \( \frac{c}{\sqrt{n}} \) are added to the empirical statistics, we can achieve the Bayes envelope (14), where the bound in equation (15) is replaced by

\[ \sqrt{\frac{12}{n} 2^k + \frac{k}{n}}. \]

All results are unchanged, except for the constant in (15) when \( x_1 \) takes values in an arbitrary (rather than binary) alphabet.
References


